Pattern Recognition: “Class” Description and Distinction

Randy Julian

Lilly Research Laboratories

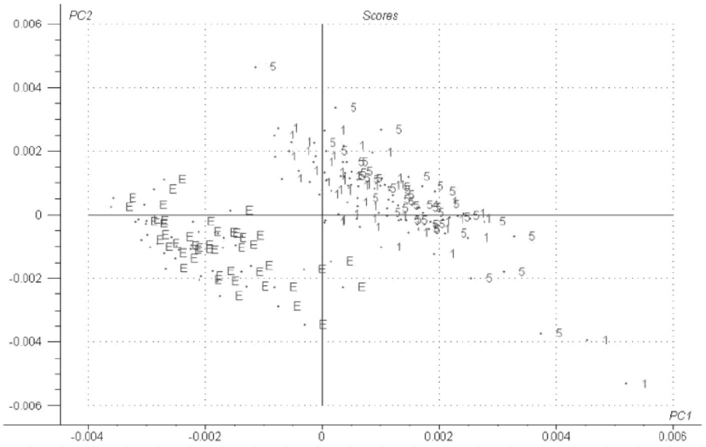
Basic concept of “recognition”

- The act of taking in raw data and taking action based on the ‘category’.
- Different from hypothesis testing:
  - Choice is always between “category” and “not category”
- Different from image processing:
  - Images go in, images come out
- Different from associative memory:
  - Data goes in, a representative pattern comes out

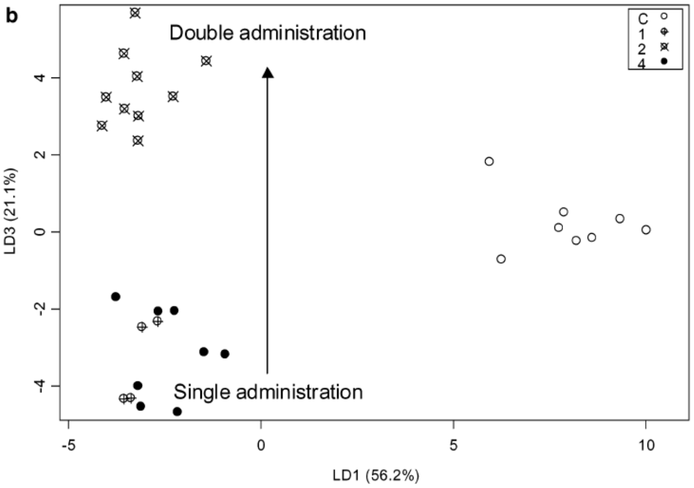
Categorization and decisions

- Pattern recognition results in the declaration of membership in a class:
  - Cannot reconstruction the pattern from the category.
  - "Feature extraction" looses some information
  - Classification looses a radical amount of information.
  - The need to make a decision between limited choices of actions forces the information loss.

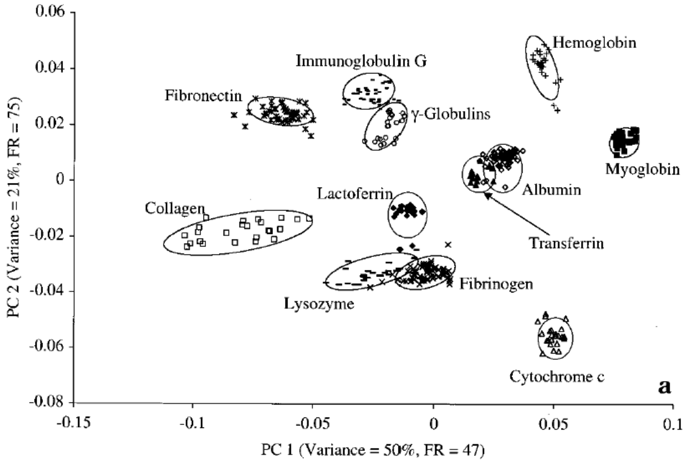
Populations: Pure vs. Contaminated



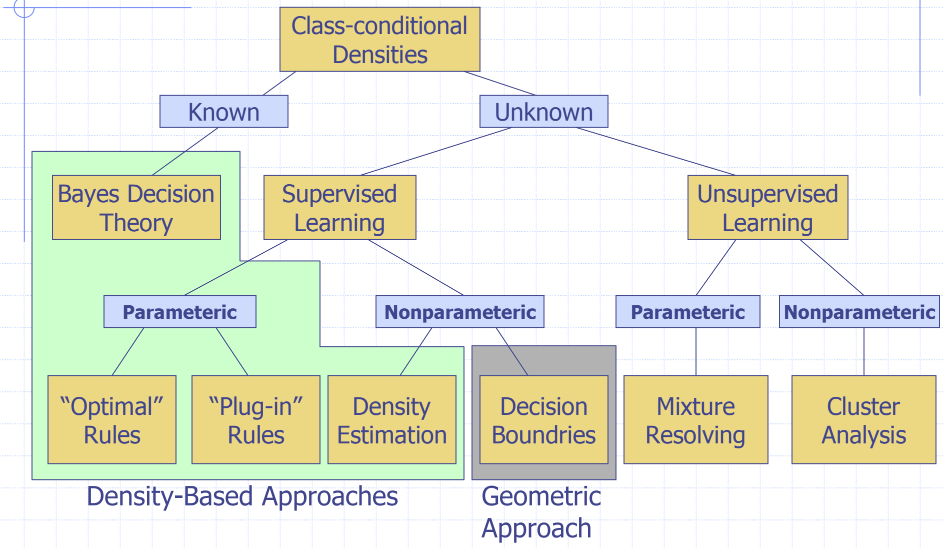
Dose-Response vs. Control



Protein Class



Various approaches



Class-conditional Densities

Known

Unknown

Bayes Decision Theory

Supervised Learning

Unsupervised Learning

Parameteric

Nonparameteric

Parameteric

Nonparameteric

"Optimal" Rules

"Plug-in" Rules

Density Estimation

Decision Boundries

Mixture Resolving

Cluster Analysis

Density-Based Approaches

Geometric Approach

Approaches to classification

- Recall: $P(\omega_i \mid x) = \dfrac{p(x \mid \omega_i) P(\omega_i)}{p(x)}$
- Density estimation:
  - Estimate density $\hat{p}_n(x \mid \omega_i)$
  - Compute $\hat{p}_n(\omega_i \mid x)$
  - Build "Plug-in" classifier $\hat{\alpha}_n(x) = \arg\max_{1 \le i \le C} \hat{p}_n(\omega_i \mid x)$
- Regression
  - Directly estimate $\hat{p}_n(\omega_i \mid x)$
- Discriminant analysis
  - Construct discriminant functions $\hat{g}_n(x,\theta)$ $\hat{\alpha}_n(x) = \arg\max_{1 \le i \le C} \hat{g}_{n,i}(x,\theta)$
  - Learn decision functions directly

$$\mathcal{D}_n \Rightarrow \hat{\alpha}(x)$$

Normal Distributions

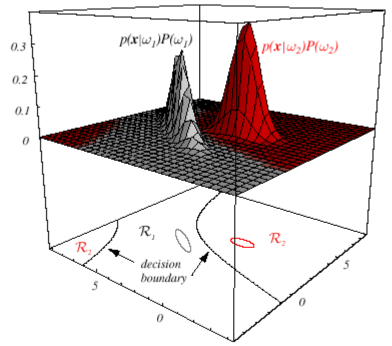


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Decision Rule

- Choose most likely class given the features measured and the likelihood of the class:

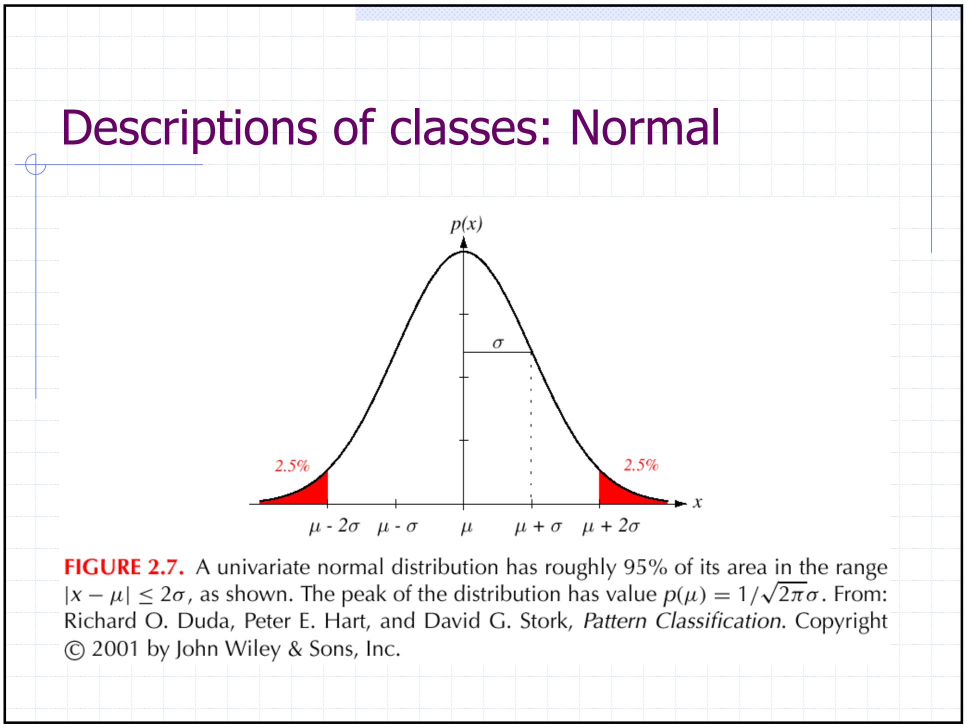
$$If\ \ P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}),\ \ choose\ \omega_1.\ \ Otherwise\ choose\ \omega_2.$$

- The scaling factor p(x) can be removed:

$$If\ \ p(\mathbf{x}\,|\,\omega_1)\ P(\omega_1)\ >\ p(\mathbf{x}\,|\,\omega_2)\ P(\omega_2),\ \ choose\ \omega_1.\ Otherwise\ choose\ \omega_2.$$

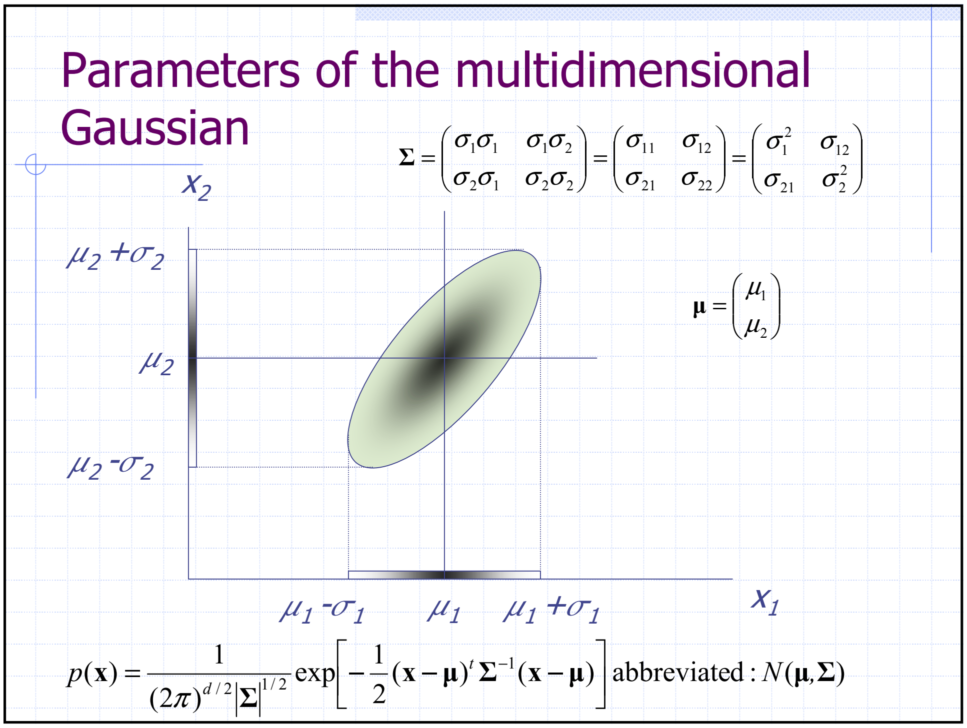
$$\begin{gathered} g_1(\mathbf{x}) = p(\mathbf{x}\,|\,\omega_1)P(\omega_1) \\ g_2(\mathbf{x}) = p(\mathbf{x}\,|\,\omega_2)P(\omega_2) \\ \text{if } g_1(\mathbf{x}) > g_2(\mathbf{x}) \text{ choose } \omega_1. \text{Otherwise choose } \omega_2. \end{gathered}$$

[1]



Descriptions of classes: Normal

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.



Parameters of the multidimensional Gaussian

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1\sigma_1 & \sigma_1\sigma_2 \\ \sigma_2\sigma_1 & \sigma_2\sigma_2 \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \text{abbreviated}: N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Quick look back at Bayes "Optimal" for Gaussians

$$g_1(\mathbf{x}) = p(\mathbf{x} \mid \omega_1) P(\omega_1)$$

$$g_2(\mathbf{x}) = p(\mathbf{x} \mid \omega_2) P(\omega_2)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \text{abbreviated} : N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Two-dimensional Gaussian

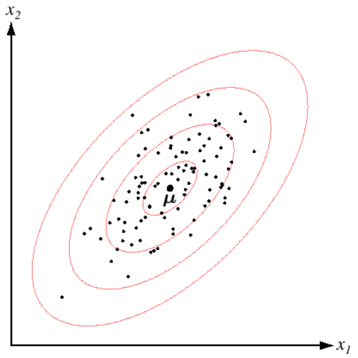
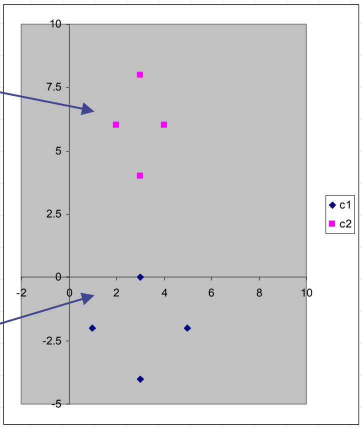


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Example

k	x1	x2
1	3	4
2	3	8
3	2	6
4	4	6
mean	3	6

k	x1	x2
1	3	0
2	1	-2
3	5	-2
4	3	-4
mean	3	-2



Some linear algebra: transpose

transpose

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \quad \text{and} \quad \mathbf{x}^t = \begin{pmatrix} x_1 & x_2 & \ldots & x_d \end{pmatrix}$$

$$\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{11} \end{pmatrix} \quad \text{and} \quad \mathbf{M}^t = \begin{pmatrix} m_{11} & m_{21} \\ m_{12} & m_{11} \end{pmatrix}$$

Some linear algebra: transpose & outer product

inner product (dot product)

$$\mathbf{x}^t\mathbf{y} = \sum_{i=1}^{d} x_i y_i = \mathbf{y}^t\mathbf{x}$$

co-linearity of two vectors

$$\cos\theta = \frac{\mathbf{x}^t\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^t\mathbf{x}}$$

Euclidian norm (length)

outer product

$$\mathbf{M} = \mathbf{x}\mathbf{y}^t = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix} = \begin{pmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_dy_1 & x_dy_2 & \cdots & x_dy_n \end{pmatrix}$$

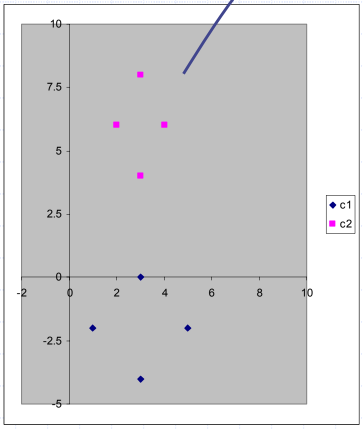
Simple estimates: Maximum Likelihood

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

simple, but biased…

Example



$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$(\mathbf{x}_k - \hat{\boldsymbol{\mu}})$$

$$(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

	mean	value	d		1	2
x1	3	3	0	1	0	0
x2	6	4	-2	2	0	4
x1	3	3	0	1	0	0
x2	6	8	2	2	0	4
x1	3	2	-1	1	1	0
x2	6	6	0	2	0	0
x1	3	4	1	1	1	0
x2	6	6	0	2	0	0
				Cov	0.5	0
					0	2
				InvCov	2	0
					0	0.5

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

$$\hat{\boldsymbol{\Sigma}}^{-1}$$

When Σ_i are not related

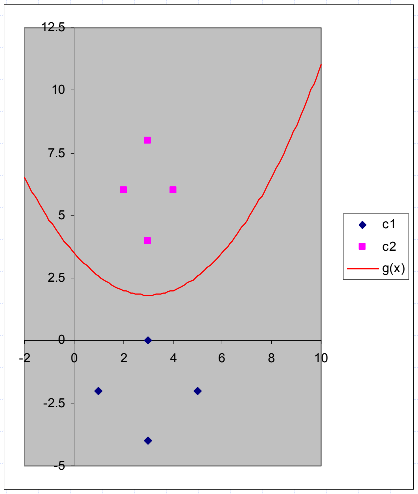
$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0}$$

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Example: conclusion



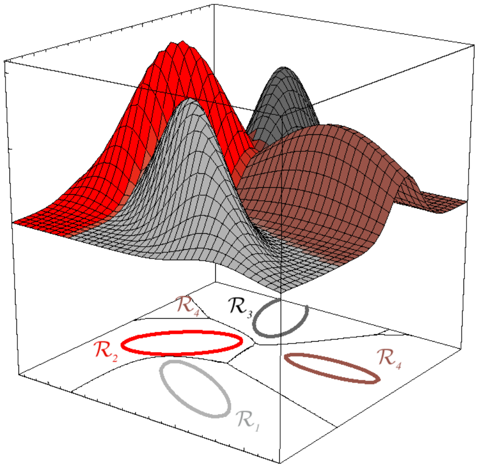
solve by setting $g_1(x)=g_2(x)$

$$P(\omega_1) = P(\omega_2) = 0.5$$

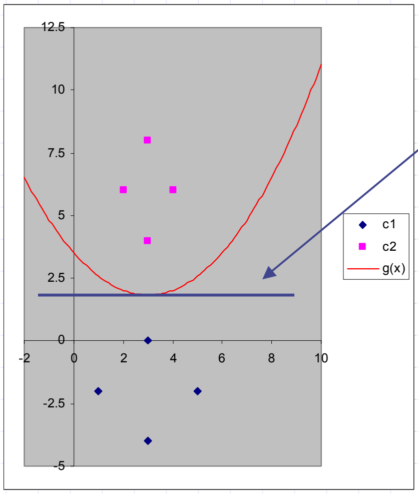
$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

the decision boundary is quadratic

Quadratic boundaries can be complex



Alternative: Least Squares



$$\hat{\mathbf{y}} = \hat{\beta}_0 + \sum_{k=1}^{n} x_k \hat{\beta}_k$$

or

$$\hat{\mathbf{y}} = \mathbf{x}^t \hat{\beta}$$

now y is class
and x is an input vector

How to 'train' a linear model

pick coefficients which minimize residual sum of squares "RSS"

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - x_i^{t}\beta\right)^2$$

$$RSS(\beta) = (\mathbf{y} - \mathbf{x}\beta)^t (\mathbf{y} - \mathbf{x}\beta)$$

differentiate w.r.t. β

$$\mathbf{x}^t(\mathbf{y} - \mathbf{x}\beta) = 0$$

$$\hat{\beta} = (\mathbf{x}^t\mathbf{x})^{-1}\mathbf{x}^t\mathbf{y}$$

set y 0 for class A,
and a 1 for class B
then if y(x) > 0.5 call it class B

$$\hat{y}(x_0) = x_0^t \hat{\beta}$$

Next time: Using R to make this easy

- How to generate populations
- How to do the linear regression on classes

References

[1] Detecting and Quantifying Sunflower Oil Adulteration in Extra Virgin Olive Oils from the Eastern Mediterranean by Visible and Near-Infrared Spectroscopy, G. Downey, P. McIntyre, A.N.Davies, *J. Agric. Food Chem.* **2002**, 50, 5520-5525

[2] Analyzing the Physiological Signature of Anabolic Steroids in Cattle Urine Using Pyrolysis/Metastable Atom Bombardment Mass Spectrometry and Pattern Recognition, M.-E. Dumas, L. Debrauwer, L. Beyet, D. Lesage, F. Andre, A. Paris, J.-C. Tabet, *Anal. Chem*. **2002**, 74, 5393-5404

[3] Interpretation of Static Time-of-Flight Secondary Ion Mass Spectra of Adsorbed Protein Films by Multivariate Pattern Recognition, M. S. Wagner, B.J. Tyler, D. G. Castner, *Anal. Chem.* **2002**, 74, 1824-1835